CEPH
Une solution de stockage distribué Open Source

Présentation XSTRA/Groupe stockage

Jeudi 05/10/2017

Sommaire

- Introduction
- Types de stockage
- Architecture réseau
- Intégration avec OpenStack
- Placement des données
  - Placement Group (PG)
  - CRUSH
  - Protection des données
  - Cache Tiering
- Performance
- Projet de déploiement à l'IPHC

CEPH

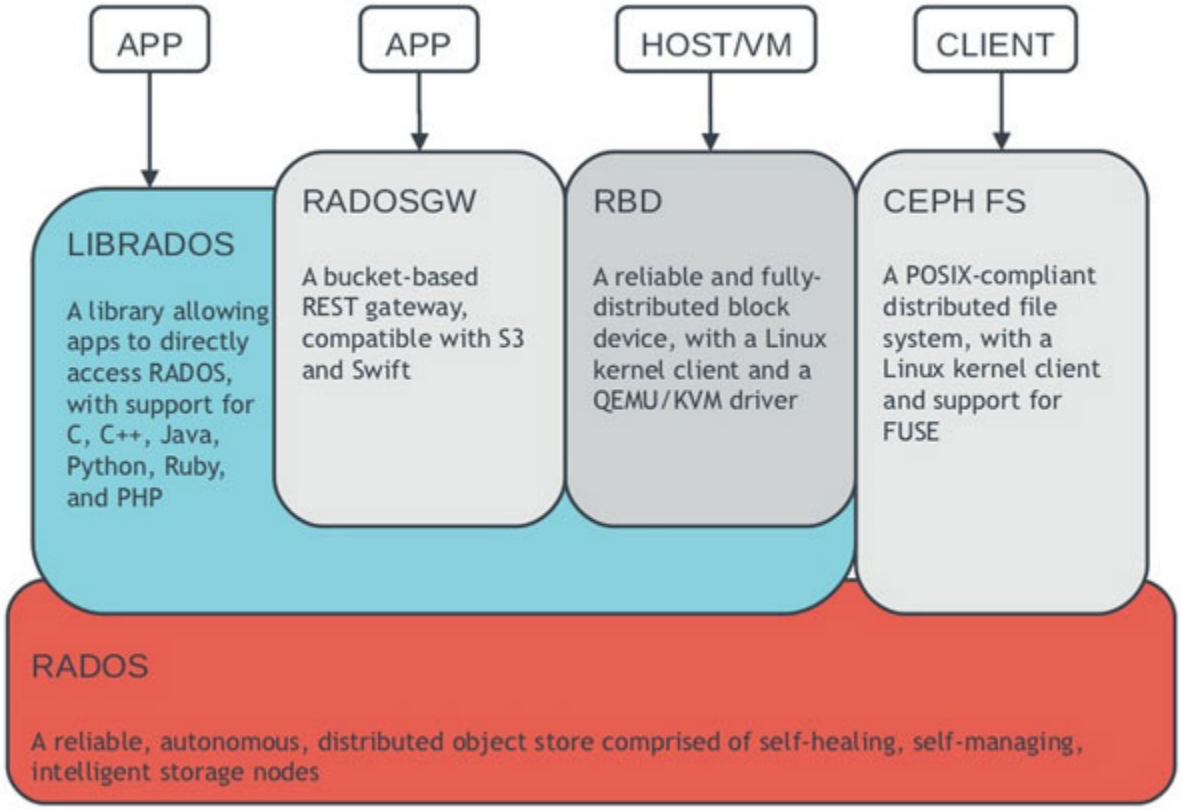
- Solution de stockage distribué
  - Pas de point unique de défaillance, les éléments sont redondés et fonctionnent en mode multi-actif
  - Extensible jusqu'à l'exaoctet
  - Conçu pour s’auto-réparer et réduire les coûts d’exploitation.
  - Offre une évolution dynamique (scale-out)
- Open Source
- Tolérance aux pannes
- Fonctionne avec du matériel standard

Historique des versions

- 12/2007 Sujet de thèse de Sage Weil
- 2011 Création de l'entreprise Inktank pour continuer le développement
- 2012 Première version LTS
- 2014 Rachat par Red Hat
- Une version LTS tous les ans
- Durée du support :
  - LTS : jusqu'à la publication de deux LTS suivantes
  - Stable: jusqu'à la version suivante

07/2012 Argonaut (v 0.48) LTS
01/2013 Bobtail (v 0.56)
05/2013 Cuttlefish (v 0.61)
08/2013 Dumpling (v0.72) LTS
11/2013 Emperor (v 0.67)
05/2014 Firefly (v0.80) LTS
10/2014 Giant (v0.87)
04/2015 Hammer (v0.94) LTS
11/2015 Infernalis (v9.2)
04/2016 Jewel (v10.2) LTS
01/2017 Kraken (v11.2)
08/2017 Luminous (v12.2) LST
xx/2017 Minic (v13.2)
xx/2018 Nautilus (v14.2) LST

CEPH 3 types de stockage



http://www.sebastien-han.fr/blog/2012/06/10/introducing-ceph-to-openstack/

Type de stockage

Stockage Object (RADOSGW)

- Espace d'adressage linéaire, accès avec un identifiant unique, support des métadonnées et des snapshots
- Exemple : stockage mail, photos, vidéos, documents en ligne type owncloud, seafile
- Pas de fonction de partage, de verrous ou d'arborescence

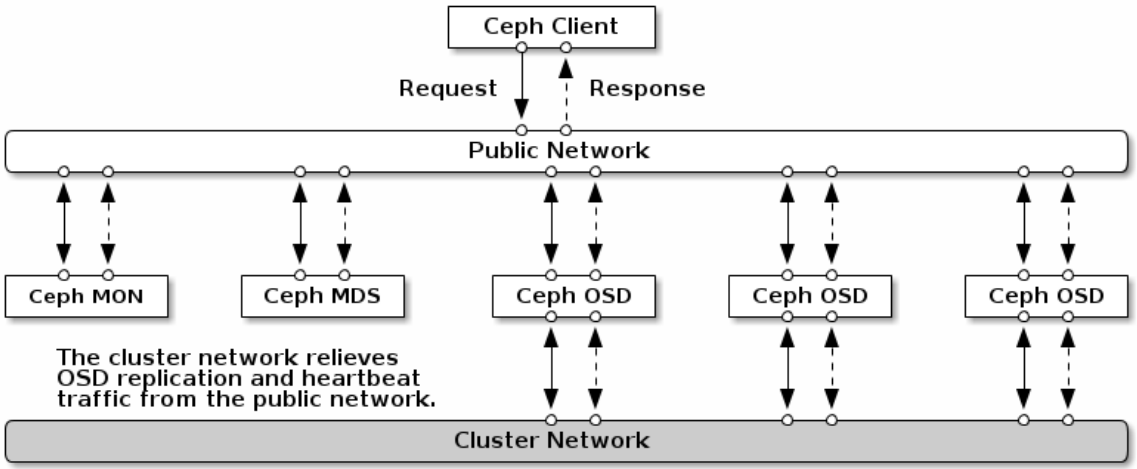
Stockage Bloc (RBD)

- Accès à un disque distant (équivalent aux disques iSCSI).
- L'espace peut être monté comme un disque local ou comme disque virtuel pour différentes machines virtuelles
- Support des snapshots, de la duplication, du thin provisionning, de la compression
- Fonction de mirorring asynchrone, cache tiering (cache rapide en lecture ou écriture), intègre une Gateway iSCSI
- Pas de gestion d'accès concurrents, pas de déduplication
- Lors de l'effacement de données, les blocs inutilisés sur le disque virtuel ne sont pas libérés automatiquement. Utiliser la commande fstrim régulièrement pour économiser de l'espace.

CEPH FS

- Accès à un stockage concurrent en mode fichiers compatible POSIX (équivalent à NFS+ACLs)
- Nécessite le service MSD (Meta Data Server)
- Support via un module kernel ou fuse
- Support snapshot par répertoire, quota par répertoire (fuse uniquement), intègre une gateway NFS
- Problème : latence avec les petits fichiers
- Non recommandé pour les disques virtuels (double écriture, sécurité)

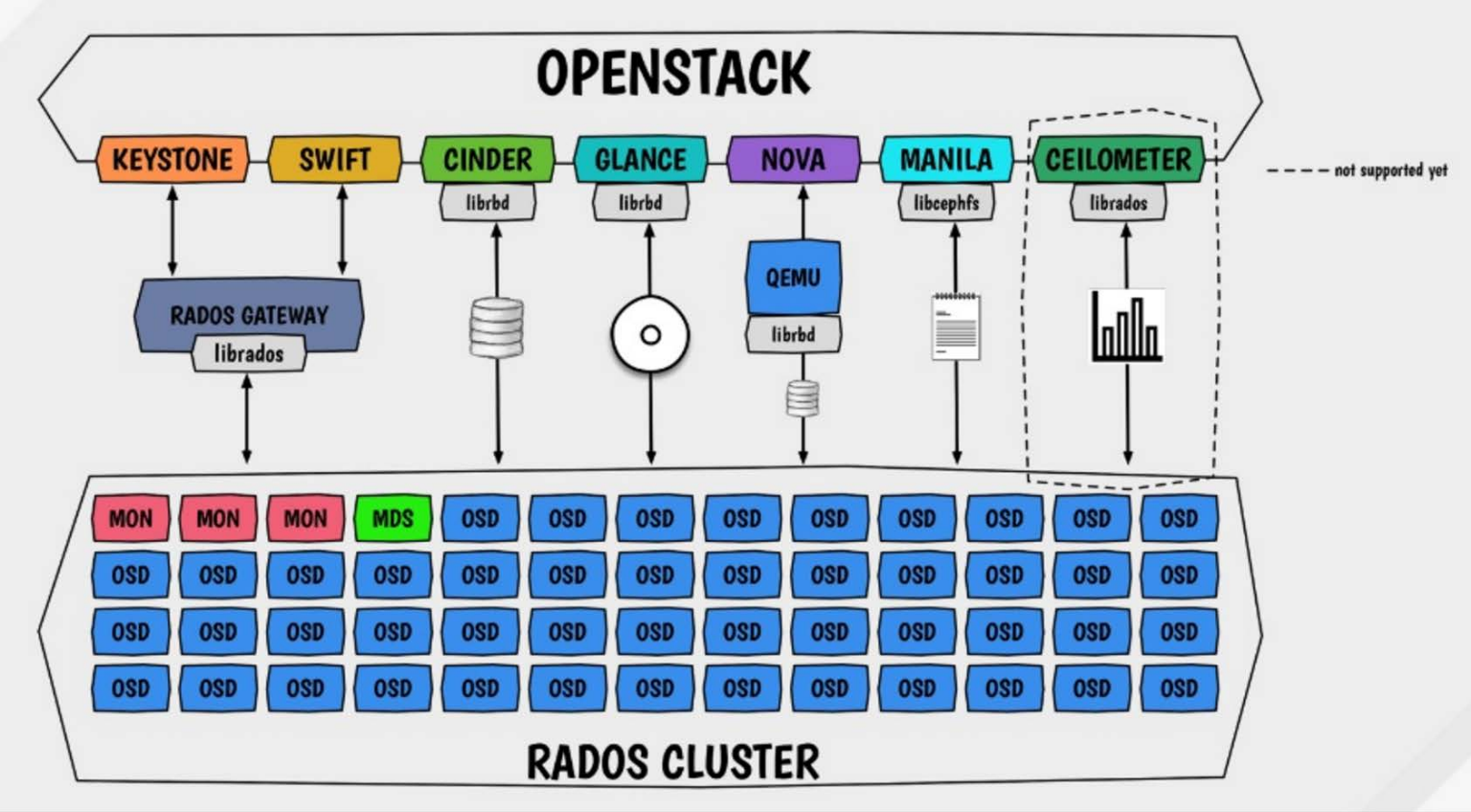
Architecture réseau



http://docs.ceph.com/docs/hammer/rados/configuration/network-config-ref/

- **MON** : service maintient une copie des cartes du cluster. Nécessite un nombre impair de services pour définir un quorum. Utiliser un serveur dédié par MON
- **MDS**(MetaDataServer) : service nécessaire uniquement avec CEPHFS. Assure l'enregistrement des métadonnées POSIX. Peux fonctionné avec les serveurs MON
- **OSD**(Object Storage Device) : service de stockage des objet. Utilise les disques locaux du serveur
- **Réseau publique** : réservé aux clients pour l'accès aux MON et aux OSD
- **Réseau cluster** : réservé pour la réplication des informations entre OSD
- Ces réseaux ne doivent pas être accessibles de l'Internet. Seul les clients doivent avoir accès au réseau public du cluster ! (attaque DOS)

Intégration avec OpenStack



https://www.sebastien-han.fr/blog/2016/05/16/The-OpenStack-Ceph-Galaxy/

Placement des Données

- Le stockage des objets se fait sur du matériel standard, sans utiliser des disques ou des contrôleurs RAID.
- Permet une meilleure évolution du matériel (disques de différentes capacités, vitesses, technologies)
- Reconstruction en parallèle =>Temps de reconstruction diminué
- Reconstruction commence sans attendre l'ajout d'un nouveau disque
- Pas besoin d'avoir des disques «hot-spares»

- Pools : Groupe logique pour stocker les objets.
  - Résilience : définie le type de réplications et le nombre de copies/réplicas d'un objet
  - Placement Group (PG) : agrégat d'objets qui permet de déterminer rapidement leurs états (accessibles, valides ou corrompus)
  - Règle Crush : détermine la distribution des objets en fonction de l'infrastructure (OSD, serveur, chassit, rack, PDU, allée, pièces, Datacenter, région)
  - Snapshots : fonction utile pour les backups ou la protection des données
  - Quota : nombres objets ou de volume maximum
  - Authentification : règles d'accès en lecture ou écriture pour les clients (service)

Placement Group (PG)

Les PG sont des fragments d'un pool logique.

Ils sont composés d'un groupe de daemons OSD qui se surveillent entre eux.

Les PG permettent :

- monitorer le placement d'objets et leurs métadonnées
- Vérifier l'interconnexion entre ces OSD (~30s)
- La gestion du placement est cher en CPU et en RAM
- Irréaliste sans PG lorsque l'on a des millions, voire milliards d'objets

Il est possible d'augmenter le nombre de PG d'un pool.

Actuellement le nombre de PG doit être un multiple de 2.

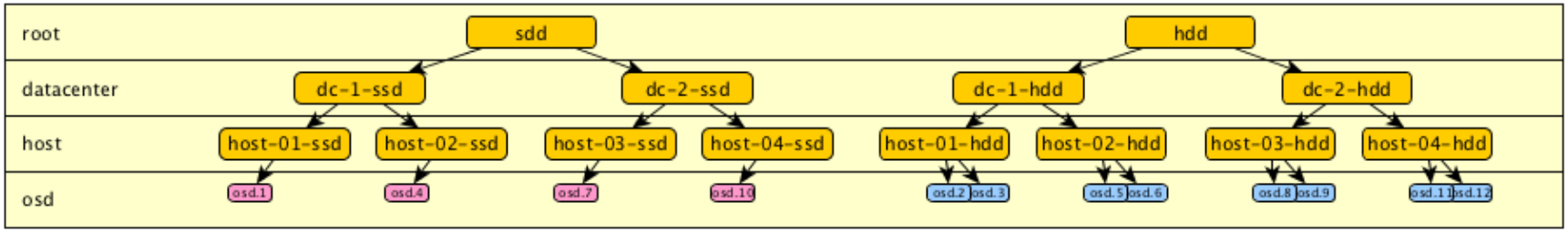
Par contre il n'est pas encore possible de diminuer le nombre de PG.

Règle à respecter :

- Entre 5 et 10 OSD PG=512, entre 10 et 50 OSD PG=4096
- Calcul du nombre de PG : PG=(OSD*100)/poolsize poolsize=nombre de réplicats

CRUSH

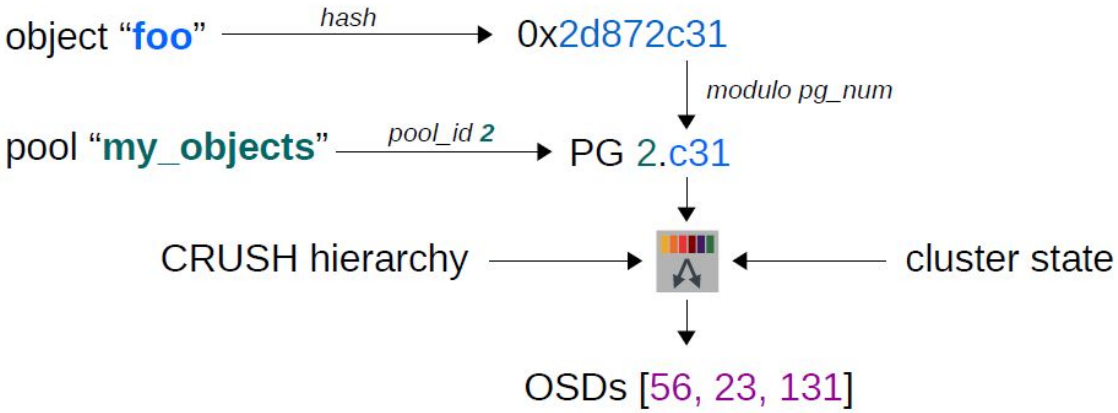
- CRUSH signifie «Controlled Replication Under Scalable Hashing»
- Algorithme de placement pseudo-aléatoire
- Calcul rapide, pas de boucle de recherche, déterministe
- Assure la distribution uniforme des informations sur les OSD
- Définit la topologie de l'infrastructure (nœuds de stockage, racks, rangées, Datacenter)
- Définit un poids à chaque OSD, son type (SATA, SAS, SSD)
- Les clients ne connaissent que les OSD, pas les serveurs ou racks…
- Exemple : matériel avec un mixte en HDD et SSD



http://cephnotes.ksperis.com/blog/2015/02/02/crushmap-example-of-a-hierarchical-cluster-map

CRUSH

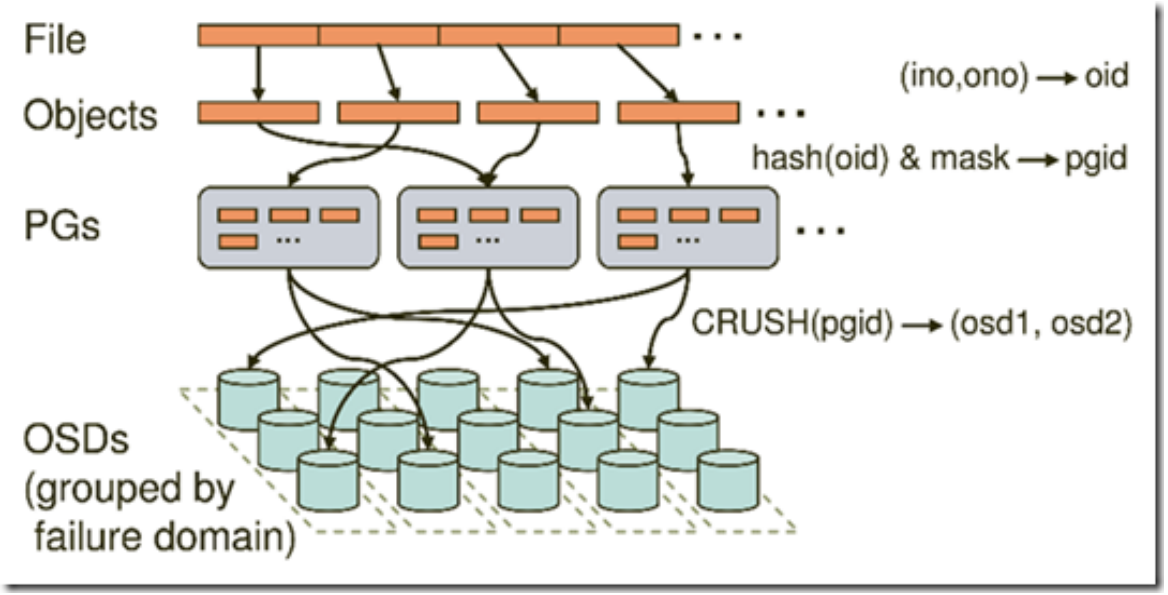
- Détermination de l'OSD en fonction du nom de la ressource



http://www.linux-mag.com/id/7744/

CRUSH

- Exemple avec un pool, une réplication de 2 et une tolérance de panne par serveur.



http://www.linux-mag.com/id/7744/

Protection des données

Depuis la version 0.80 Firefly Ceph propose deux types de protection de données :

- **Réplication** : multiples copies de la source
  - Type de réplication le plus utilisé
  - Réplication minimum par 3 pour réparer automatiquement les données détériorées
  - Reconstruction rapide sans nécessité de calcul de parité
- **Erasure-code** : codage et fragmentation de la source
  - Les données sont divisées en K fragments, puis combinées et codées avec M morceaux de données redondantes réparties sur différents OSD du cluster.
  - Protection équivalent à RAID6 : K=2, M=2 (Diminution de l'overhead à 50%)
  - Différents algorithmes disponibles : Jerasure, ISA-l, LRC, shec
  - Reconstruction coûteuse en calcul
  - Les morceaux sont repartis sur des OSD de différents serveurs => Nécessite au minimum 4 serveurs pour débuter

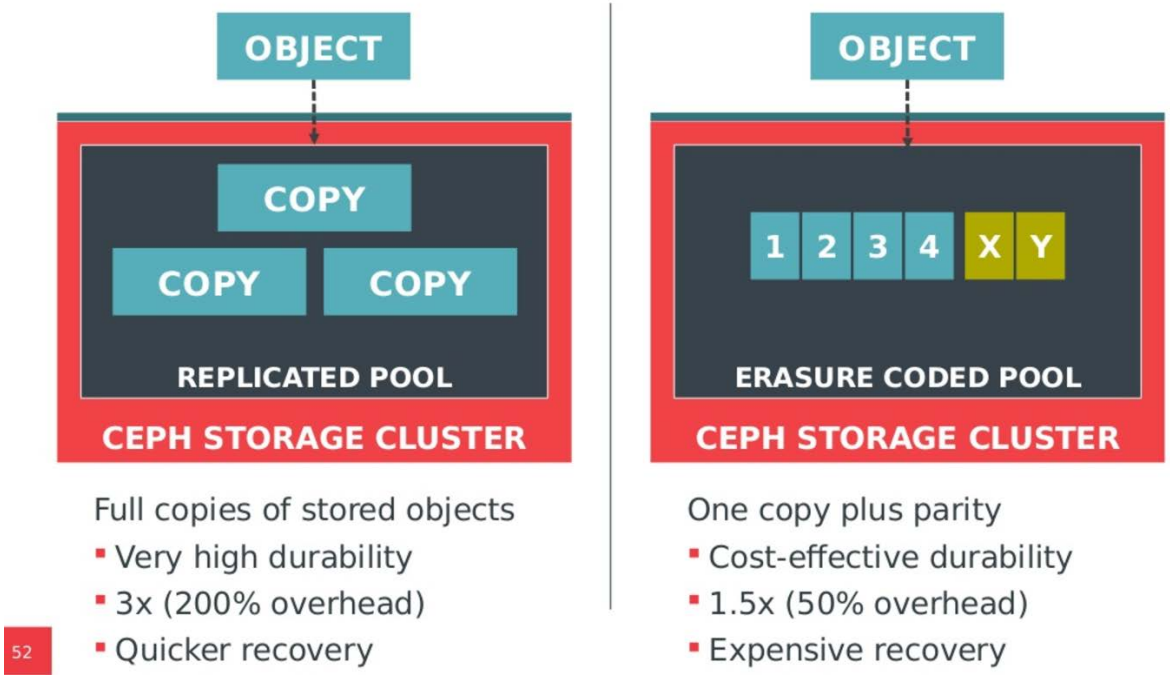
La protection des données est définit pour chaque pool

Les données sont transférées par l'OSD primaire aux réplicas

Réseau dédié à la réplication => meilleurs performances

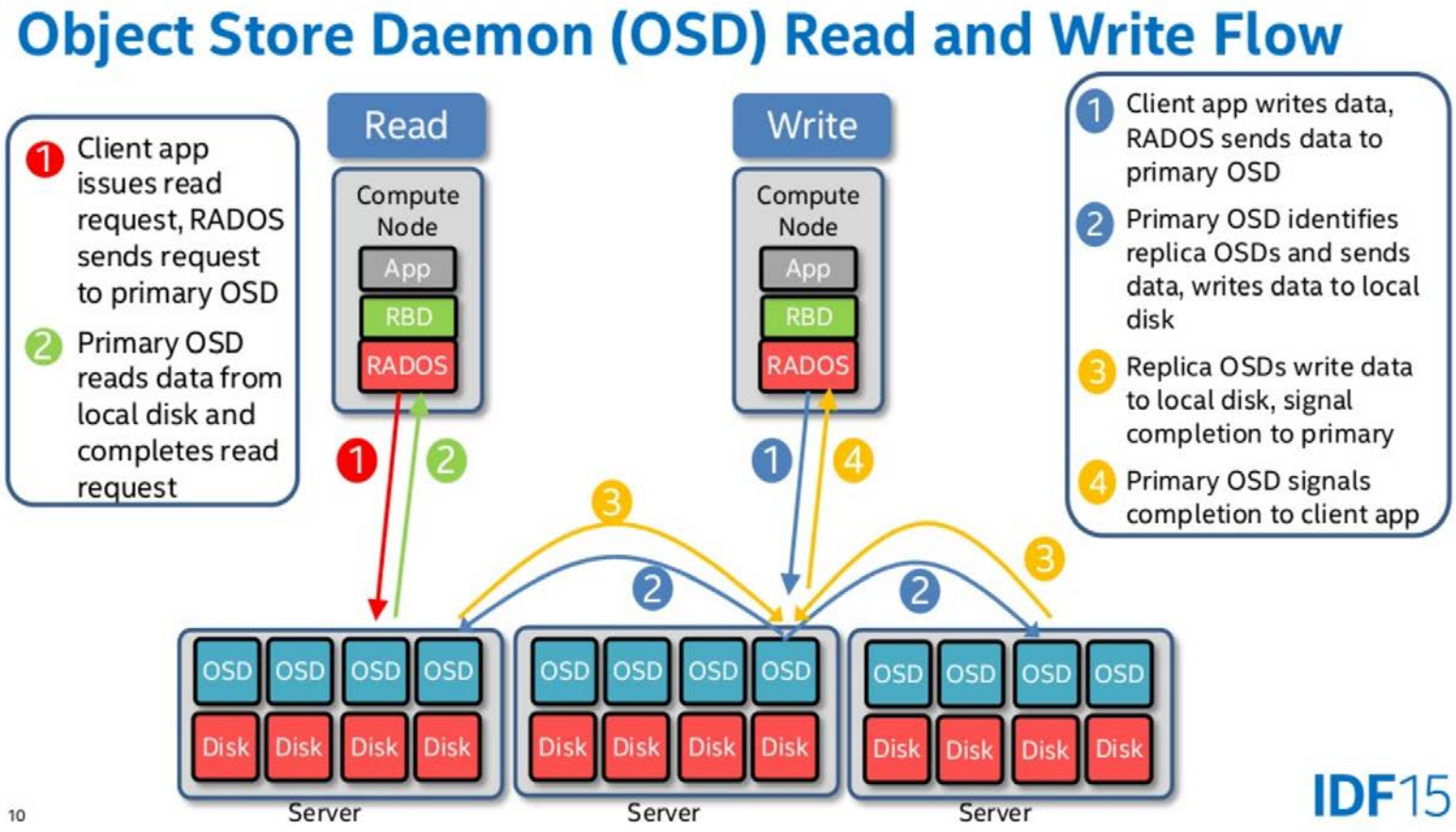
Protection des données

- Différence entre réplication et Erasure coded



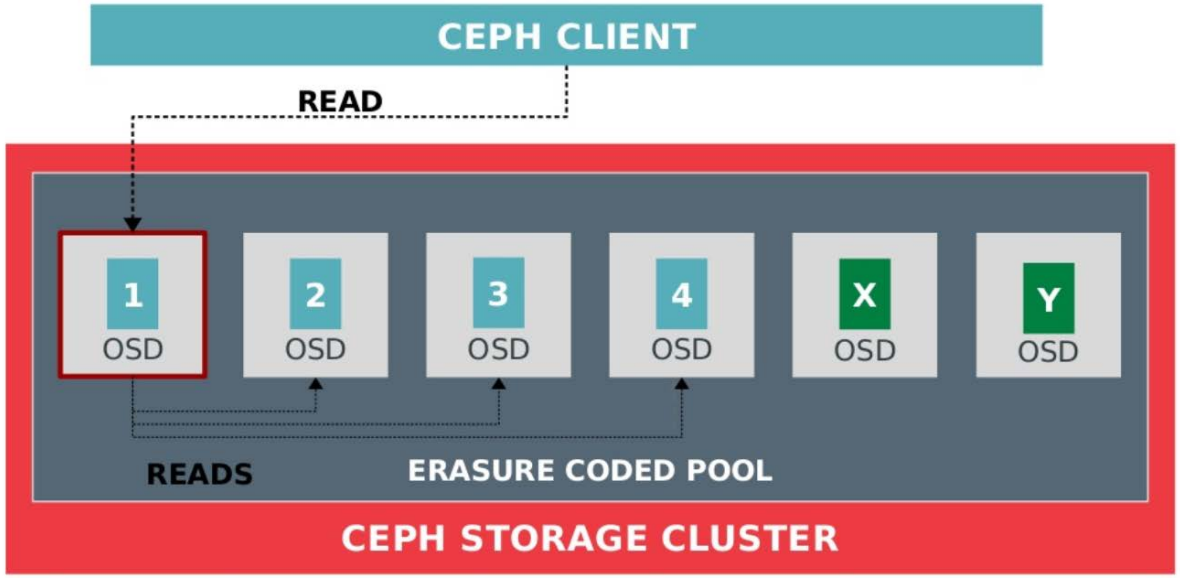
https://www.slideshare.net/sageweil1/20150222-scale-sdc-tiering-and-ec

Protection des données



Protection des données

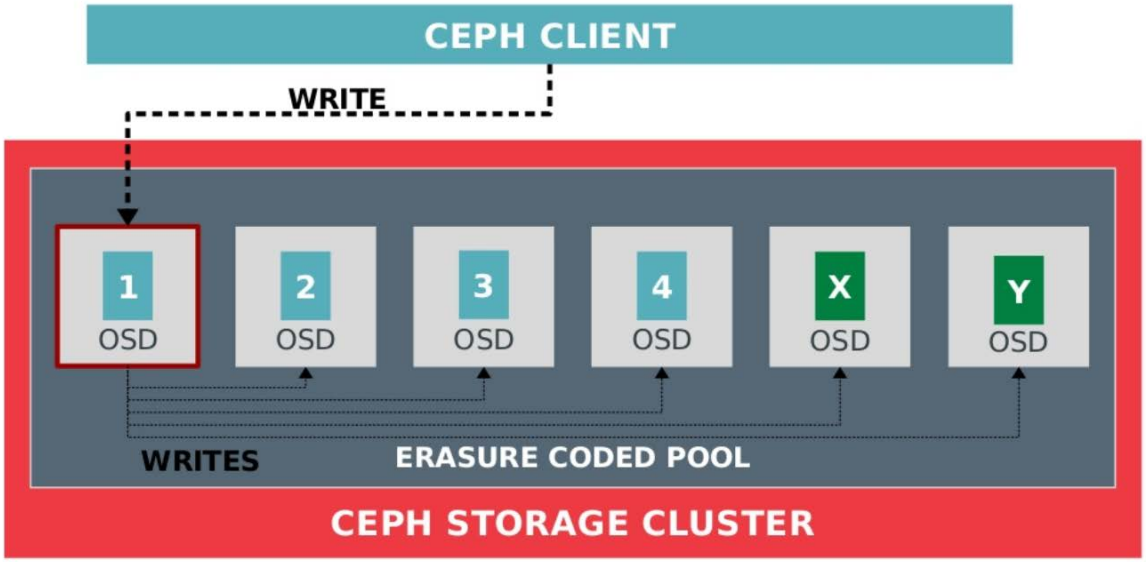
- Cycle de lecture avec l'erasure coding



https://www.slideshare.net/sageweil1/20150222-scale-sdc-tiering-and-ec

Protection des données

- Cycle d'écriture avec l'erasure coding



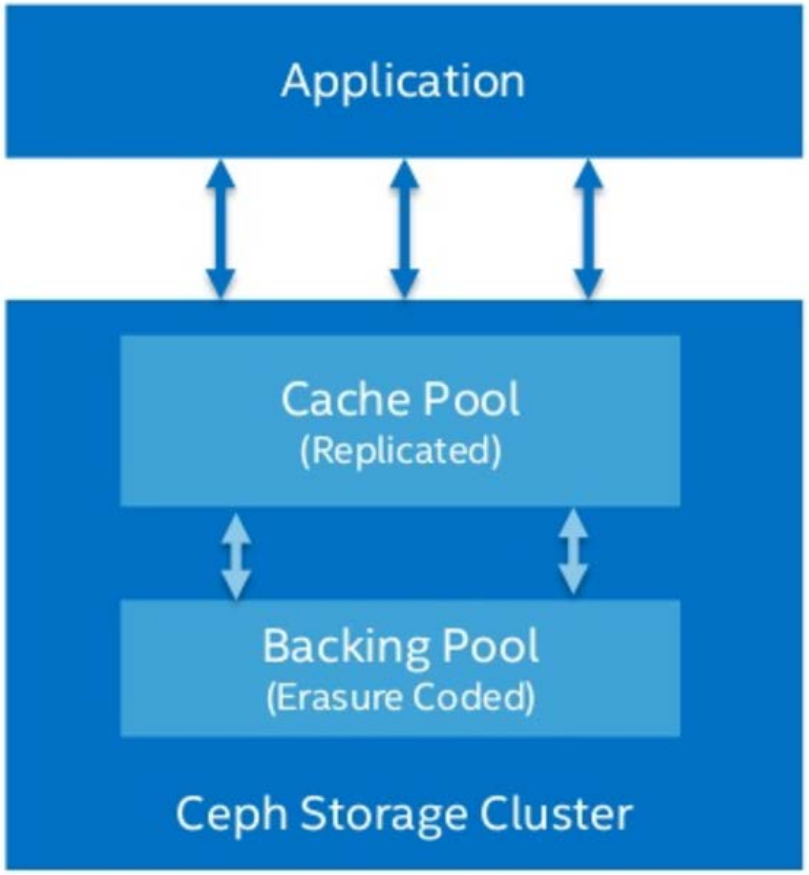
https://www.slideshare.net/sageweil1/20150222-scale-sdc-tiering-and-ec

Cache Tiering

Le Cache Tiering permet de déplacer des données «chaudes» vers les supports haute performance lorsqu'elles deviennent actives, et les données «froides» vers des supports à faible performance lorsqu'elles ne sont plus actives.

- Utilisation de disques SSD pour créer un pool de stockage rapide
- Les données sont migrées d'un pool à un autre
- Définitions de la politique pour privilégier la lecture ou l'écriture
- Définition du quota et de la température des données (nombre d'accès par heure)
- Peut être rajouté ou supprimé à chaud à tout moment sur un Pool déjà existant dans le cluster.
- Nécessaire jusqu'à Luminous pour avoir un pool RBD avec l'erasure coding

Cache Tiering



https://www.slideshare.net/LarryCover/ceph-open-source-storage-software-optimizations-on-intel-architecture-for-cloud-workloads

OSD (Object Storage Device)

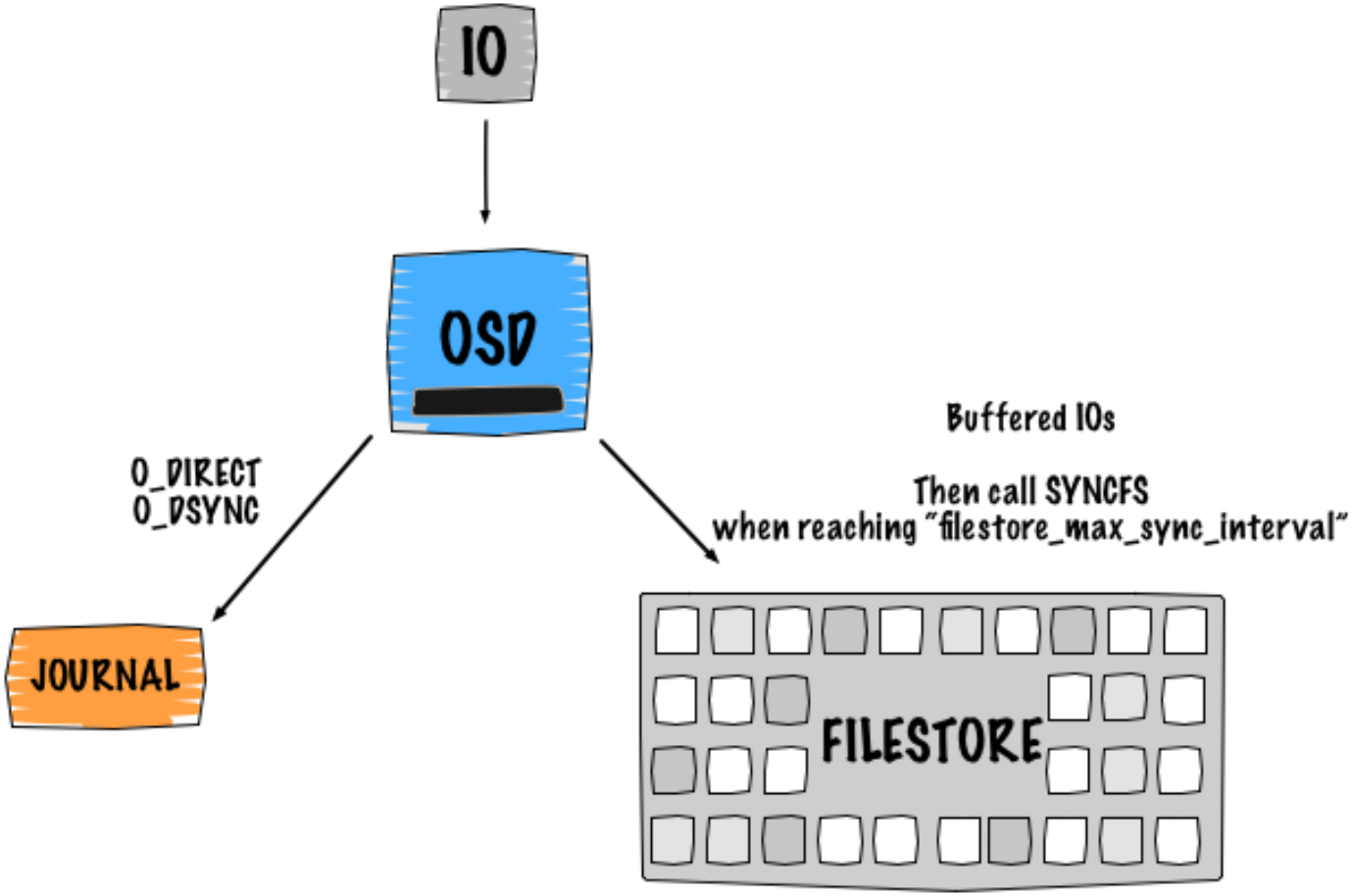
Service de stockage des objets, gère la réplication, l'intégrité des données et la récupération si nécessaire

- Les clients CEPH communiquent directement avec les OSD plutôt que par l'intermédiaire d'un serveur centralisé
- Utilisation d'un disque par service
- Evitez l'utilisation de configuration RAID ou de partitionner les disques avec plusieurs OSD
- Différents backend de stockage :
  - FileStore, BlueStore, MemStore
  - Peuvent être mixés dans un même cluster

FileStore

- Utilisation par défaut jusqu'à Luminous
- Utilisation en production (bien testé et largement utilisé)
- Utilise un journal et le système de fichiers local (XFS, BTFRS, EXT4, ZFS)
- Écriture synchrone dans le journal, puis en mode asynchrone sur le disque=>Provoque une double écriture !
- Optimisation : possibilité de déplacer le journal sur un disque séparé (SSD) pour augmenter les performances.
- Attention : La perte du disque dédié au journal provoque l'arrêt de tous les OSD concernés. Il est plutôt conseillé d'utiliser le Cache Tiering

FileStore

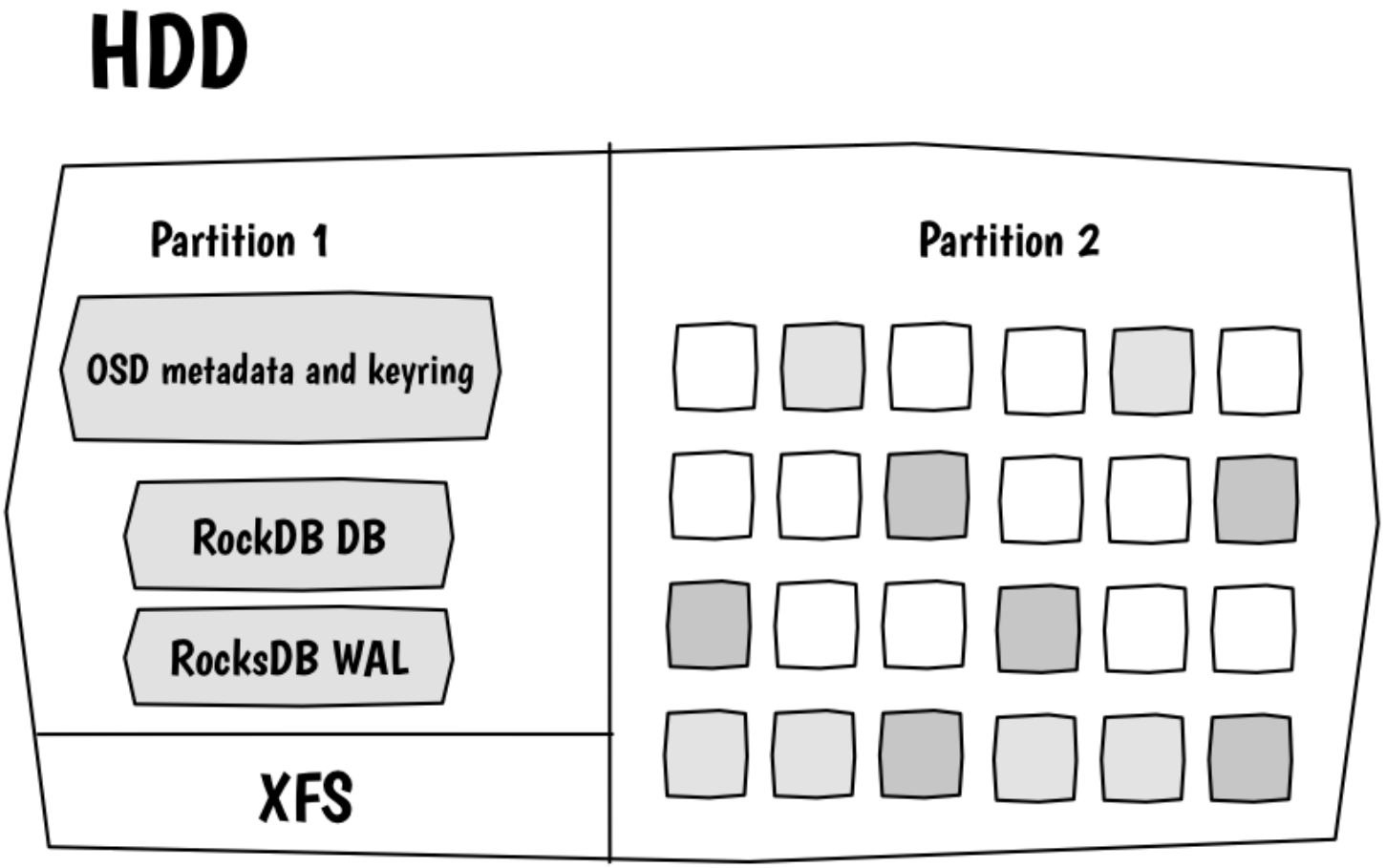


https://www.sebastien-han.fr/blog/2013/12/02/ceph-performance-interesting-things-going-on/

BlueStore

- Version stable depuis Luminus. Utilisé par défaut.
- Écrit directement les données sur le disque sans passer par un journal
- Utilisation système de fichier simplifié (bluefs)
- Plus de doubles écritures
- rockDB pour l'enregistrement des métadonnées (énumération plus rapide)
- Augmentation de la taille des caches par OSD (1GB SATA et 3GB SSD par défaut)
- Gain en vitesse écriture X2 sur les disques SATA et plus sur les SSD
- Checksums : vérification de la cohérence à chaque lecture. Permet de diminuer les processus en tâche de fond pour vérifier l'état des PG (scrubing)
- Activation de la compression : lz4, snappy, zlib
- Procédure bluestore-migration pour passer de FileStore à BlueStore http://docs.ceph.com/docs/master/rados/operations/bluestore-migration/

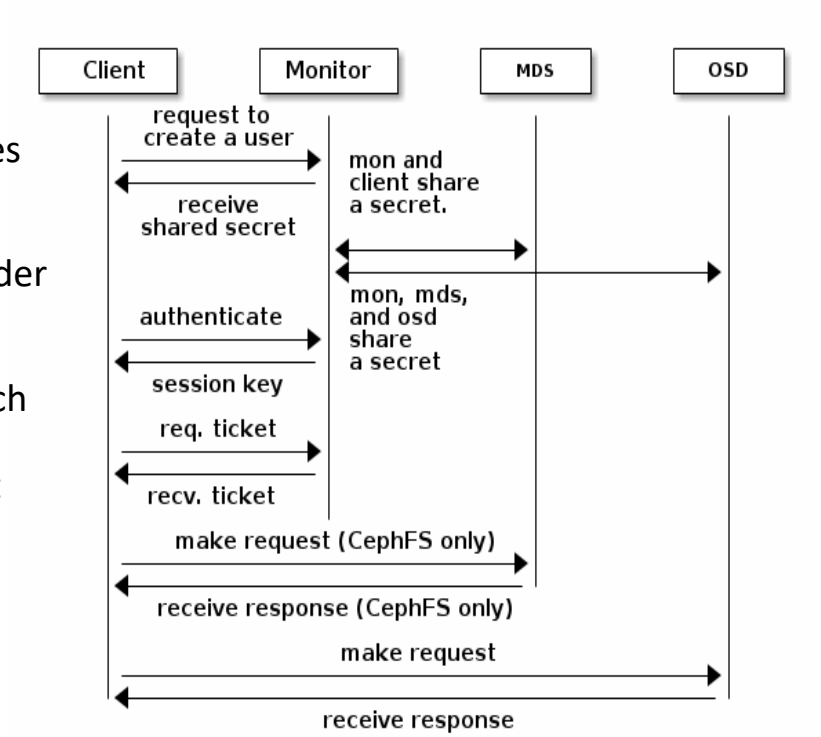
BlueStore



https://www.sebastien-han.fr/blog/2016/03/21/ceph-a-new-store-is-coming/

Authentification : CephX

- les MON connaissent
  - les clefs de tout le monde
  - les autorisations
- Le client s'authentifie via les MON et demande l'accès aux serveurs
- le MON génère un «secret partagé» limité dans le temps pour faire ses demandes de tickets pour accéder aux serveurs
- RAPPEL : un client écrit uniquement dans l'OSD primaire qui s'occupe de la réplication (ou du dispatch en EC)
- Protection contre l'attaque « man in the middle », et contre les attaques par re-jeu des paquets
- Fonctionnement proche de kerberos
- Limite:
  - N'est pas destinée à l'authentification des humains
  - Pas de chiffrement des transferts

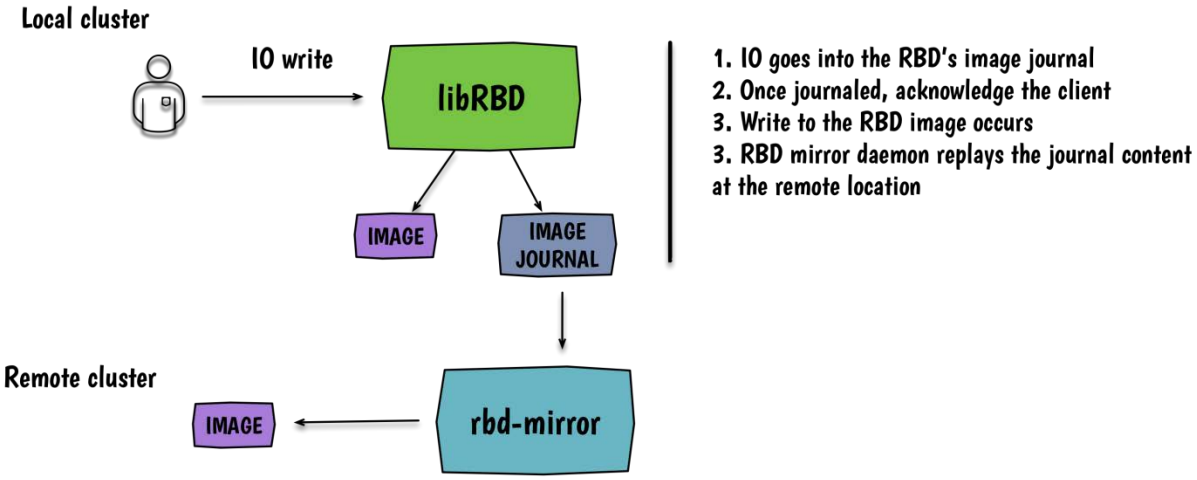


http://docs.ceph.com/docs/firefly/rados/operations/auth-intro/

RBD-mirror

- Nouveau service de réplication asynchrone entre cluster
- Disponible depuis la version Jewel
- Multiples services depuis Luminous
- Réplication active / passive ou active / active
- Utilisation d'un journal pour enregistrer les transactions à synchroniser.
- Peut synchroniser tout un pool ou sélectionner les images à synchroniser
- Définir l'image principale (promote / demote)
- Les clusters doivent avoir des noms différents
- Manque : définition QoS, temps de rétention d'une image supprimée, délais de réplication

RBD-mirror



https://www.sebastien-han.fr/blog/2016/03/28/ceph-jewel-preview-ceph-rbd-mirroring/

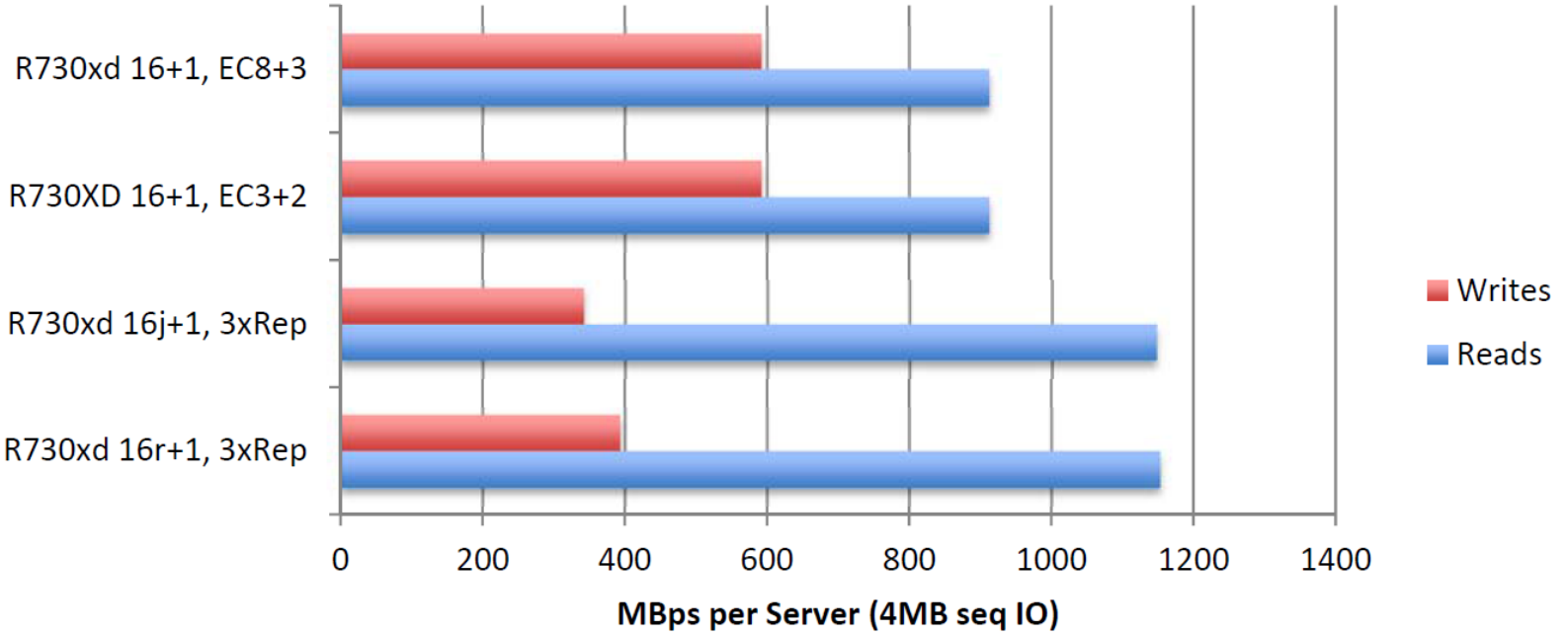
Configuration nécessaire

- MON, MSD
  - RAM: 1GB par service , CPU: > 4 cœurs
- OSD
  - RAM : 1GB par service , CPU > 2 Cœurs
  - Pendant les reconstructions , le besoin de RAM augmente de ~1Go pour 1To de stockage à reconstruire
  - Moins de 20 disques par serveur
  - SSD : utilisé pour les journaux. Ne pas dépasser 5 journaux par SSD
- Contrôleur de disques
  - Activer la configuration en mode RAID0 pour les disques SAS ou SATA, et JBOD pour les disques SSD
- Réseaux
  - 2*10GB/s ou 1*FB à 40BG/s
- Documentation
  - Dell PowerEdge Performance and Sizing guide for CEPH Storage http://en.community.dell.com/techcenter/cloud/m/dell_cloud_resources/20442913
  - http://docs.ceph.com/docs/master/start/hardware-recommendations/

Comparaison de performances

Réplication vs. Erasure-coding

- Configuration : disques 4To SAS dans 15 R730xd
- J:JBOD r:RAID0

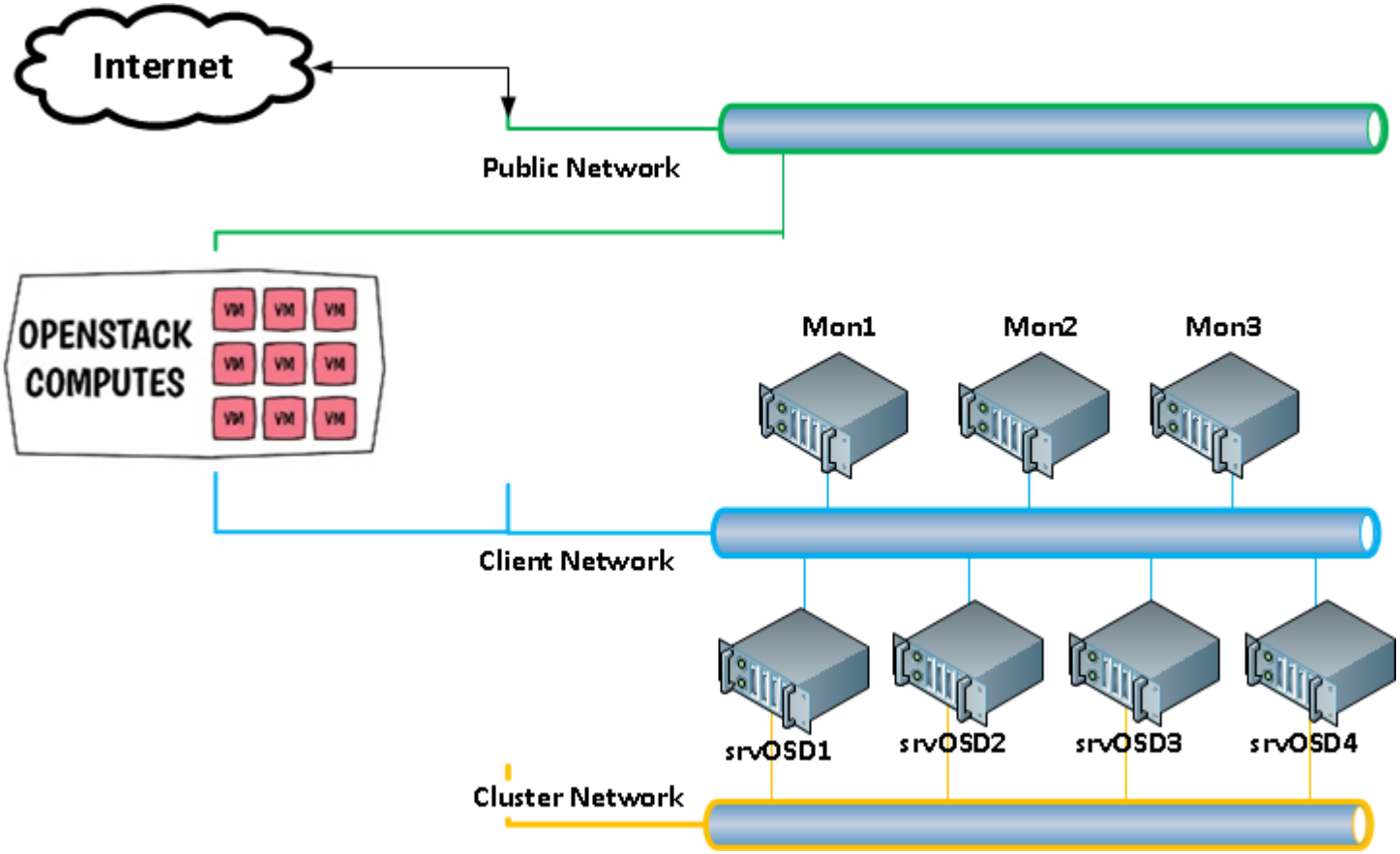


Dell PowerEdge Performance and Sizing guide for CEPH Storage

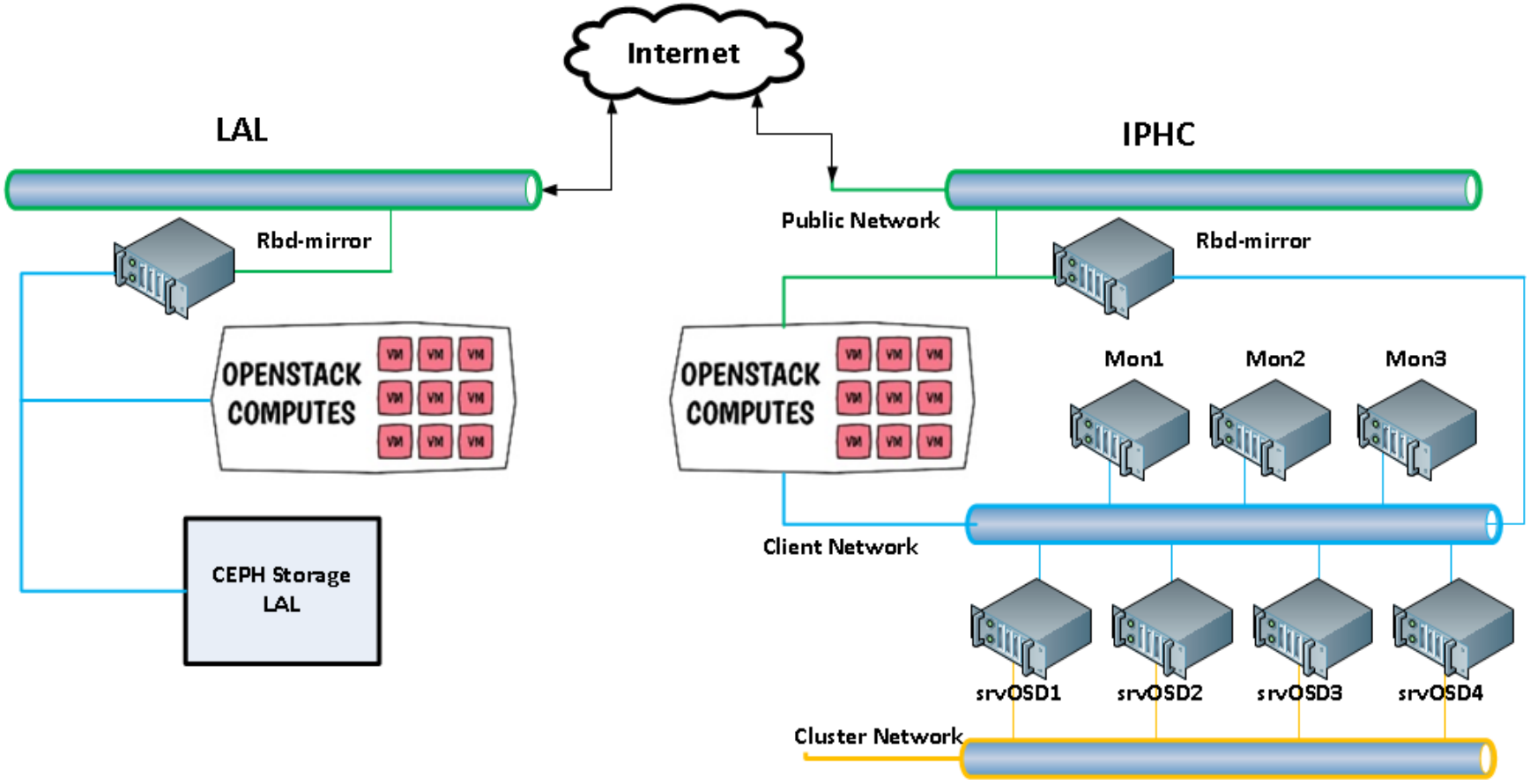
Projet de déploiement à l'IPHC

- Service de stockage RBD pour OpenStack
  - 3 MON : M630 8 cœurs 16Go de RAM
  - 48 OSD : 4*730xd 10 cœurs 64Go de RAM
    - (8To*12*4)=>384To brute
    - 128To réplication*3 ou 256To Erasure Coding
  - 1 RBD-mirror (réplication avec le LAL)
- Financement: France Grilles et CPER

Projet de déploiement à l'IPHC



Projet de déploiement à l'IPHC



Annexes

- CEPH http://docs.ceph.com/docs/master/
- http://www.sebastien-han.fr/blog/
- Dell PowerEdge Performance and Sizing guide for CEPH Storage http://en.community.dell.com/techcenter/cloud/m/dell_cloud_resources/20442913
- ANF CNRS 2017 CEPH https://groupes.renater.fr/wiki/ceph/public/formation2017